REPORT DOCUMENTATION PAGE

Form Approved
OMB NO. 0704-0188

1. AGENCY USE ONLY (Leave Blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	17, May 03	Final Report (01, March 03 - 28, Feb 03)

4. TITLE AND SUBTITLE
"Detection of Biological Warfare Pathogens by Rare Event Imaging"

5. FUNDING NUMBERS
Contract # DAAD-19-01-1-0321

6. AUTHOR(S)
Lan Bo Chen, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Dana-Farber Cancer Institute

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U. S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSORING / MONITORING AGENCY REPORT NUMBER
42301.1-LS

11. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision, unless so designated by other documentation.

12 a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

12 b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

The standard methods for the detection and identification of pathogens require that either a sufficient amount of such pathogens are present or that these pathogens are grown under selective conditions. These procedures are time consuming and inadequate in many situations. The objective of this project was to establish a rapid and extremely sensitive method to detect and identify BW pathogens in our environment and in human body fluids using the revolutionary approach of Rare Event Imaging. During the grant period different immunofluorescence and fluorescence in situ hybridization based protocols were developed for the sensitive and specific detection of pathogens. The Rare Event Imaging System (REIS) was adopted to be able to detect the fluorescently labeled rare cells fast and reliably. Simultaneous enumeration of multiple pathogens was achieved with the REIS. The utility of the fully automated REIS approach in clinical diagnostics was shown with using the Cytomegalovirus antigenemia assay.

14. SUBJECT TERMS
Fluorescence microscopy
Cytomegalovirus (CMV)
Immuncytochemistry
Immunofluorescence
Rare event detection
Image analysis

15. NUMBER OF PAGES
22

16. PRICE CODE

17. SECURITY CLASSIFICATION OR REPORT	18. SECURITY CLASSIFICATION ON THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

NSN 7540-01-280-5500

Standard Form 298 (Rev.2-89)
Prescribed by ANSI Std. 239-18
298-102

Enclosure 1

20030604 107

Table of Contents:

Problem studied:

The standard methods for the detection and identification of pathogens require that either a sufficient amount of such pathogens are present or that these pathogens are grown under selective conditions. These procedures are time consuming and inadequate in many situations. The objective of this project was to establish a rapid and extremely sensitive method to detect and identify BW pathogens in our environment and in human body fluids using the revolutionary approach of Rare Event Imaging.

The Rare Event Imaging System (REIS) is an automated, epifluorescence microscope–based image analysis platform developed in this laboratory for the sensitive detection of cancer cells in the peripheral blood. During the grant period we developed procedures for the fluorescent labeling of microorganisms and adopted the Rare Event technology to be able to detect them.

One of the major challenges of our approach was to accelerate the analysis process to a level that is adequate for most purposes. We addressed this issue from several perspectives. First, we developed various in situ hybridization-based protocols that yielded high signal/noise ratios. Second, we adapted the system to work at the lowest possible microscope magnification while still allowing reliable identification of microorganisms. Third, we developed image analysis algorithms to better distinguish between true positive and false positive hits.

Specific Aims:

- Development of methods for specific visualization of pathogens
- Optimization of procedures for analysis in the REIS
- Adopt the extremely sensitive and rapid automated detection system, the Rare Event Imaging System (REIS), for pathogen detection
- Simultaneous enumeration and identification of pathogens in water/air samples and human body fluids

Project Milestones:

1. Adaptation and optimization of basic procedures and development of probes
2. Development of the Multiplex procedure
3. Detection, enumeration and identification of microorganisms in environmental samples and in human body fluids
4. Adaptation of the REIS for the detection of microorganisms
5. Proof of principle: Rapid and automated detection of CMV infected leukocytes in the peripheral blood using the REIS approach

1. *Adaptation and optimization of basic procedures and development of probes*

First we tested and optimized various slide coating procedures using gram-stain and acridinorange labels to find those that could best be coupled to REIS analysis. We determined that the bacteria adhere firmly to gelatin-coated slides and to specially coated and charged adhesive slides (Marienfeld GmbH, Germany). Appropriate cell deposition conditions have been worked out to ensure maximal recovery of bacteria.

Once the plating procedures were developed we tested various commercially available monoclonal and polyclonal antibodies for the labeling of bacteria. Using E. coli and S. epidermidis as models we established procedures for the immuncytochemical labeling of specific bacterial species. The fixation and labeling conditions were optimized in these experiments to achieve bright fluorescent signals with minimal background staining.

In parallel to developing immuncytochemical labeling procedures we acquired fluorescein labeled oligonucleotides and experimented with *in-situ* hybridization based detection of bacteria. We determined appropriate in-situ hybridization conditions for gram-positive and gram-negative bacteria and reproducibly achieved bright fluorescent signals in these preparations. To ensure hybridization specificity each probe was tested in the respective target strain as well as in related microbial species together with a universal bacterial probe (EUB) and a reverse complementary probe (NUB).

FISH based detection was achieved for the following pathogens:

Gram negative	Gram positive
E. Coli	S. Epidermidis
S. Maltophilia	B. Subtilis
B. Cepacia	

In further experiments we decided to use *in situ* hybridization based detection of pathogens because we encountered the following problems with immuncytochemistry:

- production of antibodies is time-consuming
- limited number of targets can be detect simultaneously using indirect IF
- low commercial availability
- sensitivity and specificity problems

2. *Development of the Multiplex procedure*

We performed mixing experiments with pre-calculated proportions of different bacterial species. Probes for simultaneous and specific detection of bacteria species were designed and coupled to distinct fluorochromes. The reliability of the multiplex procedure was validated by performing object counts with the REIS at the appropriate channel of labeling. In order to accurately count fluorescent-labeled bacteria we established optimal imaging conditions for each objective magnification (4x-100x) and determined the morphometric counting criteria (threshold, area, major axis, minor axis, length, width, etc.) for each species. Raw and corrected automated counts were closely correlated with the initial bacteria dilutions and manual counts. To mimic the detection of `rare` bacteria, low bacteria of one kind was mixed into bacteria within other kinds and then detected by counting the specific probe's signal by the REIS.

3. *Detection, enumeration and identification of microorganisms in environmental samples and in human body fluids*

After developing the multiplex procedure we focused on the evaluation of the performance of the REIS technology on environmental water and human peripheral blood samples. By definition rare event detection involves the analysis of a large sample size in order to identify a single positive event. Therefore, we introduced a membrane filtration step into the analysis process and enriched pathogens on the surface of black polycarbonate filters from these suspensions.

To avoid immediate clogging of the narrow (0.2 _m) pore-size filters that were required to retain all bacteria from the specimen, we first needed to work out appropriate sample preparation conditions for human peripheral blood and environmental water samples. The major steps of the sample preparation involved enzymatic digestion, lyses of human cells and carefully controlled flow-rate.

The enriched bacteria were examined by two different methods:

1. Membrane transfer technique
2. Direct epifluorescent membrane technique

The membrane transfer technique relies on the transfer of pathogens from the surface of the membrane filters to conventional glass microscope slides and subsequent labeling and analysis of pathogens on the slides. The advantage of this approach is that several filters can be transferred to a single slide and thus a new filter can be used if it is clogged. However, a major drawback is that there is significant cell loss during the transfer process. We extensively experimented and refined different transferring conditions but were not able to recover >75% of the spiked bacteria from water samples.

The direct epiflurescent membrane technique does not involve any transferring steps and the pathogens are examined directly on the surface where they are concentrated. However, in contrast to the membrane transfer technique, the sensitivity of the direct epifluorescent membrane method is limited to the volume that can be filtered through the area of the membrane. To avoid the embedding of cells in the membrane pores and thus a shifting out of the optical plane during scanning, the direct epiflurescent membrane protocol was optimized to retain bacteria at the membrane's surface.

The classical direct epiflurescent membrane technique does not permit the detection/enumberation of specific bacterial species; therefore, we had to develop immuncytochemical and FISH protocols to specifically visualize bacterial species on the membrane filters. After modifying our existing labeling protocols we were able to successfully detect bacterial cells directly on the membrane filters with both unique sequence-specific FISH probes and mono and polyclonal antibodies. With both procedures we observed that neither the labeled objects nor the background had a consistent brightness throughout the whole membrane filters. The image analysis component of the REIS was updated with a dynamic threshold determination algorithm that allowed accurate segmentation of labeled objects irrespective of the high-noise and high-background environment of the membrane filters.

4. *Adaptation of the REIS for the detection of microorganisms*

The REIS was originally developed for the detection of human cancer cells and thus during the grant period certain technical modifications had to be made to adopt the system to detect and enumerate microbes. The major modifications are listed bellow:

Problems	**Description of problems**	**Current features of the REIS**
Fixed scanning magnification	Digital imaging of bacteria staining does not allow image analysis based detection at very low (4x, 10x) magnifications because of the low signal intensity and pixel resolution	Scanning can be run on any magnification that the microscope is equipped with. Separate capture parameters can be assigned to each magnification.

Fixed scanning area	Border images have a higher level of autofluorescence and therefore the capture parameters are not functioning properly on them.	A flexible scanning area can be defined and saved. A positional filtration algorithm discards objects that lie outside the predefined scanning pattern.
Lack of autoexposure	The integration times needed to be determined manually for each channel. A minimal variance in the staining intensity of the preparations resulted in unreliable target identification.	An autoexposure algorithm was developed and implemented. Automatically determining the proper integration ensures that the maximum amount of image information for the processing necessary to determine true positives from false positives is gathered.
Lack of autofocusing	Due to the small size of bacteria and the high magnification (20x-60x) that is required to image them out of focus images were frequently encountered. Identification and quantification of bacterial cells based on morphometric object analysis strongly depends on the availability of high-quality images.	A fast and reliable autofocus module was introduced into the REIS. In addition, to reduce the total scanning time a focal plane prediction feature was designed and implemented.
Single channel detection	The program could capture images of positives in 1, 2 or 3 color channels but were not able to determine if a positive is single, double or triple labeled.	A multi-count feature was designed and implemented in the REIS. The feature allows keeping track of objects at multiple channels and match signals if they are derived from the same object.

Total cell count	The total cell counting algorithm that was developed to count human cells were not accurate to determine bacteria numbers.	New filtering algorithms to determine cell boundaries and cell group boundaries from the images were developed. Even on clumpy bacteria preparations total cell numbers could be counted reliably with the new algorithm.
Fixed image thresholding	Especially on environmental and human blood samples we observed a certain inconsistency in the overall intensity of the images. A fixed image threshold resulted in loosing objects or detecting too many false positives.	A dynamic intensity histogram analysis-based thresholding method was designed. The algorithm reliably segmented objects irrespective of the background intensity.
Lack of image parameter measurements	To be able to discover new capture parameters the image analysis measurements need to be recorded in an organized way.	A centroid data table displays all the image analysis measurements and flags at which step of filtration the objects were lost. This data can be uploaded into a PostgreSQL database and be used for capture parameter discovery.

5. Proof of principle: Rapid and automated detection of CMV infected leukocytes in the peripheral blood with using the REIS approach

INTRODUCTION

To test the practicability of the REIS technology for the rapid diagnosis of infectious disease in the clinical diagnostic setting we initiated collaboration with a group at Yale University School of Medicine (New Haven, CT) specializing in detecting CMV antigenemia in clinical patient samples.

CMV infection may result in significant morbidity and mortality in transplant, AIDS and cancer patients. A manual fluorescent microscopy based antigenemia assay has been widely used for rapid diagnosis and monitoring of CMV infection.[1,2,3] In high-risk transplant recipients early detection of low-level antigenemia has an utmost importance.[4, 5] Thus automated analysis of clinical specimens would be highly desirable not just to replace the long and tedious manual slide reading but also to improve the accuracy and reproducibility of the assay.

We hypothesized that the immunoflurescently labeled CMV-infected cell nuclei can be automatically distinguished from non-labeled cells, non-specifically labeled cytoplasms of eosinophils and other debris on the slides by their morphometric characteristics. The aims of the present work were to develop an automated approach to identify the classifiers that effectively discriminate with a high degree of certainty true from false positive cells and to evaluate these classifiers against manual analysis.

MATERIALS AND METHODS:

Cell preparation:
The CMV infected leukocytes were detected on cytospin preparations of peripheral blood polymorphonuclear cells with the CMV Brite Turbo Kit. (Biotest, USA). The kit contains a cocktail of two monoclonal mouse antibodies (C10/C11) that targets the CMV lower matrix phosphoprotein (pp65), an early antigen in virus replication that is abundantly present in antigen positive cells. The kit was used as per the manufacturer instruction except that the secondary antibody of the kit was replaced with an anti-mouse Alexa-488 conjugated antibody (Molecular Probes, OR) and for nuclei counterstain 0.5 _g/ml DAPI was used. The slides were mounted with the ProLong mounting Media (Molecular Probes, OR). The brightly positive cells showed homogenous green polylobate nuclear staining when examined under green fluorescence emission.

REIS analysis and morphologic evaluations:
The stained slides were scanned by the REIS using a 4x magnification lens (Nikon Plan Apo, numerical aperture=0.2) and the FITC filter cube (Chroma, VT) with a single band-pass epi illuimnation. The cytofuge concentrated the cells into a spot on the slides that could be covered by 30 (5x6) 4x images. Digital images were grabbed with a high-resolution CCD camera (Cooke, Sensicam) using 2.1s integration time and the gain setting of 3. Other components of the hardware and the scanning software have been described previously.[6,7] The objects are segmented from their background by a dynamic thresholding method based on image intensity histogram analysis. Cytomorphometry measurements were performed on the segmented objects with the Image-Pro software (MediaCybernetics, CA), which is using 49 different measurements to characterize the objects.

Three different methods were developed to use in conjunction with the Rare Event Imaging System (REIS) to identify the most suitable parameters capable of detecting CMV-infected cells.

1. The '**Range Prediction**' method discovers morphological parameters by statistically analyzing raw data to discover the upper and lower cutoff points of those parameters which best separate CMV positive cells (i.e. true positives) from non-labeled cells and debris (i.e. false positives). The Range Prediction method is effective in creating parameter ranges (upper and lower) that are likely to incorporate all of the true positives cells.

2. The '**Decision Tree**' method discovers morphological parameters by analyzing raw data using the C5.0 induction algorithm, a data mining technique developed by Ross Quinlan and based on his C4.5 algorithm.[8] A decision/classification tree uses a set of rules to predict what category a target variable falls into. The set of rules generated by C5.0 are developed using splitting rules to recursively partition a set of training data to the point where an accurate classification can be made. The rules themselves consist of a decision tree model of REIS parameters and a breakpoint measurement that determines which branch of the tree to travel down when that parameter is above or below the given value. Each branch of a tree ultimately culminates in a leaf node that consists of the final decision about that case, as well as an accuracy ratio that states how often the cases in the training set were misclassified in that node.

3. The '**Combination**' method combines both the Range Prediction and Decision Tree methods to isolate true positives from false positives. As a result of the cooperative use of the two methods, the total number of items need to be reviewed at the end of a REIS scan is filtered twice, resulting in a shorter total scanning time.

Data collection:

The training data set was obtained by manually locating true positive cells on 10 different slides from 10 CMV-infected patients. After a positive cell was manually found, the area of the slide where it was located was digitally photographed using the REIS. The images were analyzed using ImagePro and the morphometric measurement data generated from all objects were output to a table along with the proper classification. An example of the data table is shown in Table 1.

Table 1: Morphometric data collection (actual table consists of all 49 measurements):

obj_num	lomag_area	lomag_aspect	lomag_axis_major	lomag_radius	true_positive
1	16	1.046573	5.372485	3.365728	0
2	13	1.896172	5.885183	2.939244	0
3	46	1.356388	9.339807	4.943009	1
4	70	4.652574	22.26914	11.12938	0
5	168	1.343718	17.07281	8.933394	0
6	38	6.407779	19.99971	9.947926	1
7	252	2.770644	33.95271	20.48535	0
8	14	3.545789	8.020912	3.252954	1
9	20	9.475558	15.66096	7.221158	0

After all the positives were photographed on the 10 slides, the data was compiled into one set of raw data resulting in 135 true positive cells and 3153 false positive objects (Table 2).

Table 2: Training data set of 10 patients` samples

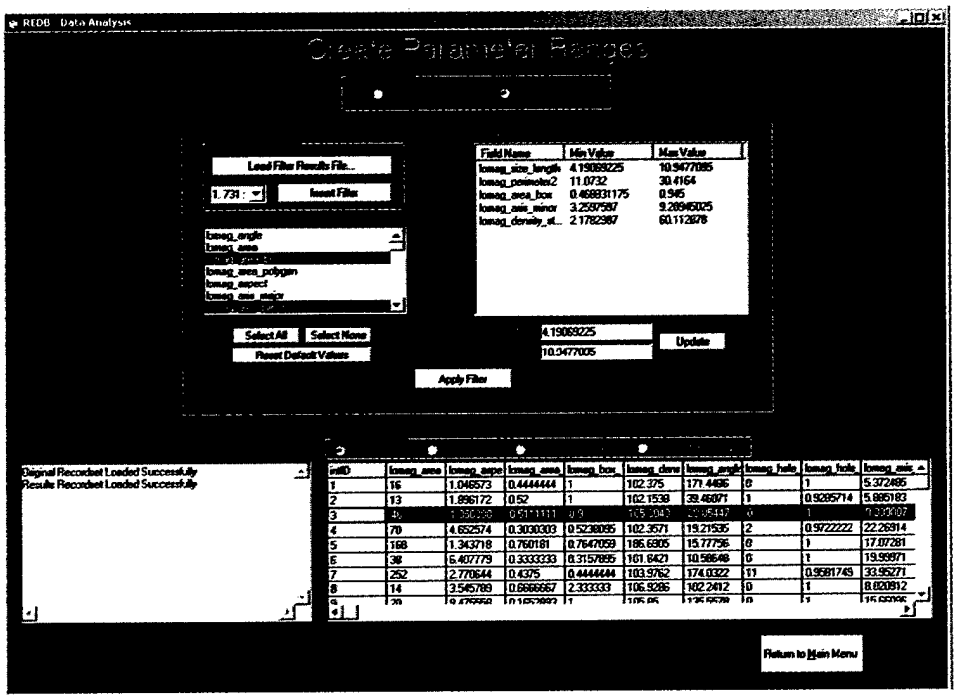
Slide Number	Number of True Positives	Number of False Positives
1	10	643
2	1	51
3	2	75
4	7	204
5	12	467
6	2	36
7	4	94
8	34	531
9	33	644
10	30	408
Total	**135**	**3153**

Data mining tools:
A custom-built application using the Visual Basic programming language was developed to implement the Range Prediction, Decision Tree, and Combination methods.

Range Prediction
There were three main components of the Range Prediction method:

1. Raw data analysis
2. Optimal parameter-combination discovery
3. Virtual testing

A screenshot of the Range Prediction application is pictured below:



1. Raw data analysis

Raw data analysis proceeded by calculating and comparing, per each parameter, the medians, standard deviations and upper and low ranges of true and false positives. After the analysis was done, parameters were automatically classified as optimal or non-optimal. An optimal parameter was one that, within a calculated range, kept every true positive yet removed false positives. Conversely, a non-optimal parameter was one that lost any true positive. After analyzing the raw data set, 44 of the 49 morphological measurements were determined to be optimal parameters. Once all the parameters were classified, a filtering algorithm iteratively applied the parameter ranges to the raw data set and ranked the optimal parameters according to both the number of overall and

unique false positives they removed. This information was used later in the optimal parameter-combination discovery phase. The algorithm also produced a best-case-scenario number that represented, given the sample set of data and a 100% true-positive retention rate, how many false positives were unable to be removed using all 44 optimal parameters. The best-case-scenario number from our raw data was 425, equivalent to an 86.5% reduction of all false positives. The large reduction in the number of false positives proved our hypothesis that CMV positive cells can in fact be distinguished from false positive objects by cytomorphometric parameters.

2. Optimal parameter-combination discovery

While the 44 optimal parameters efficiently reduced the number of false positives, using that many imaging parameters in the REIS was impractical for two main reasons. First, the more parameters used to filter incoming data, the greater the chance was that true-positives would be mistakenly classified as negatives. Second, if a false-negative was created, an abundance of parameters made it hard to identify, analyze, and fix the cause for the misclassification, making it difficult to improve for future uses. Therefore, the aim of optimal parameter-combination discovery was to create a combination of parameters that produced a number close to the best-case-scenario number of 425 while reducing the number of optimal parameters from 44.

The optimal parameter-combination tool could be customized in three ways:

a. Inputting the desired number of final parameters in each combination

Due to the computing power necessary to calculate all the various combinations of parameters, the running time of the algorithm grew exponentially as the size of the parameter combination increased. Therefore, while limiting the number of false-positives was a goal, the length of time to discover that number was also a factor when choosing which combinations could be analyzed. Under these conditions, we compared the efficiency of parameter combinations that varied in length from 1 to 7 on a Pentium 4 personal computer with 1 GB of memory. (Table 3)

Table 3: Effect of the size of parameter combinations to remove false positives and their total running time.

Combination Size	False-Positives Remaining	Total Running Time (Minutes)
1	1696	0.00007
2	1062	0.0158
3	879	0.221
4	778	2.26
5	731	18.1
6	692	117.65
7	681	638.68

Since the five-parameter bundle provided a balance point between a low false positive count and a low total running time of the algorithm it was chosen to use in further experiments.

b. Inputting the total number of combination results to return

There were multiple parameter combinations that removed the same or very close number of false-positives. Some of these combinations were known to carry parameters that possess a high risk of creating false-negatives. The optimal parameter-combination tool allowed for selecting how many top combinations of parameters the algorithm should return. The most optimal parameter bundle could then be selected from these top hits. Using the five parameter-bundle, the software returned the top 15 parameter combinations along with how many false-positives were created by each one. (Table 4)

Table 4: Example of Top Parameter Combinations.

Combination 1	**Parameter Name**	**Min**	**Max**
False Remaining: 731	lomag_size_length	4.19069225	10.94771
True Removed: 0	lomag_perimeter2	11.0732	30.4164
	lomag_area_box	0.468831175	0.945
	lomag_axis_minor	3.2597597	9.20945
	lomag_density_std_dev	2.1782987	60.11288
Combination 2	**Parameter Name**	**Min**	**Max**
False Remaining: 732	lomag_perimeter2	11.0732	30.4164
True Removed: 0	lomag_feret_max	4.24693605	10.96226
	lomag_area_box	0.468831175	0.945
	lomag_axis_minor	3.2597597	9.20945
	lomag_density_std_dev	2.1782987	60.11288
Combination 3	**Parameter Name**	**Min**	**Max**
False Remaining: 736	lomag_feret_max	4.24693605	10.96226
True Removed: 0	lomag_area_box	0.468831175	0.945
	lomag_axis_minor	3.2597597	9.20945
	lomag_density_std_dev	2.1782987	60.11288
	lomag_box_x_y	0.54285717	1.75

c. Modifying the parameter ranges

The sample size of the training data set and the possible variance of digital cytomorphomteric measurements allowed for the possibility that the ranges retrieved from the raw data might not completely cover all possible values of true positives. Thus, if ranges were to be predicted strictly based on the training data, a large possibility existed to loose true positives on a new, independent data set. To compensate for this, a certain percentage could be added to the top and bottom of each optimal parameter range. This also allowed for the fact that some optimal parameters demonstrated a very slim difference with respect to the values of the true positive range and the false positive range. Adding a

percentage to the parameter ranges caused these ranges to overlap, effectively removing from the results those parameters that had a very thin separation between true and false positives.

3. Virtual testing

Virtual testing was the capability to apply the results from the optimal parameter-combination discovery to any set of data produced from the ImagePro image analysis. The goal of such testing was to examine how the numbers of true and false positives were affected by altering the ranges returned by the optimal parameter discovery. It also allowed for testing and validating the performance of a newly discovered parameter combination on an independent data set. Using virtual testing we identified which parameters returned by the optimal parameter-combination discovery method were most sensitive to change, as those parameters were excluded from the final set because they ran a high risk of removing true positives.

For example, one of the top parameter ranges returned by the optimal parameter-combination algorithm included a parameter called lomag_box_x_y. Based on the raw data, it was determined that any object that had a lomag_box_x_y measurement not between 0.54285717 and 1.75 was to be labeled a negative. While such a classification held true for the 135 positives in the raw data set, based on test data analysis done using the virtual testing tool, it was determined that this parameter was not optimal as true positives were lost due to the narrow range of acceptance.

As a result of the optimal parameter-combination discovery and virtual testing tools, the raw data analysis was completed with the following results. A final parameter combination size of 5 parameters was chosen and the top 15 resulting combinations were produced and tested using an added percentage value of 5%. The top ranked combination resulted in a total of 731 false positives remaining and consisted of the following parameters: lomag_size_length, lomag_perimeter2, lomag_area_box, lomag_axis_minor and lomag_density_std_dev. While 731 false positives represented an increase of 72% over the best-case-scenario number of 425, it still signified a 76.8% reduction in the overall number of false positives while using 88.6% fewer parameters.

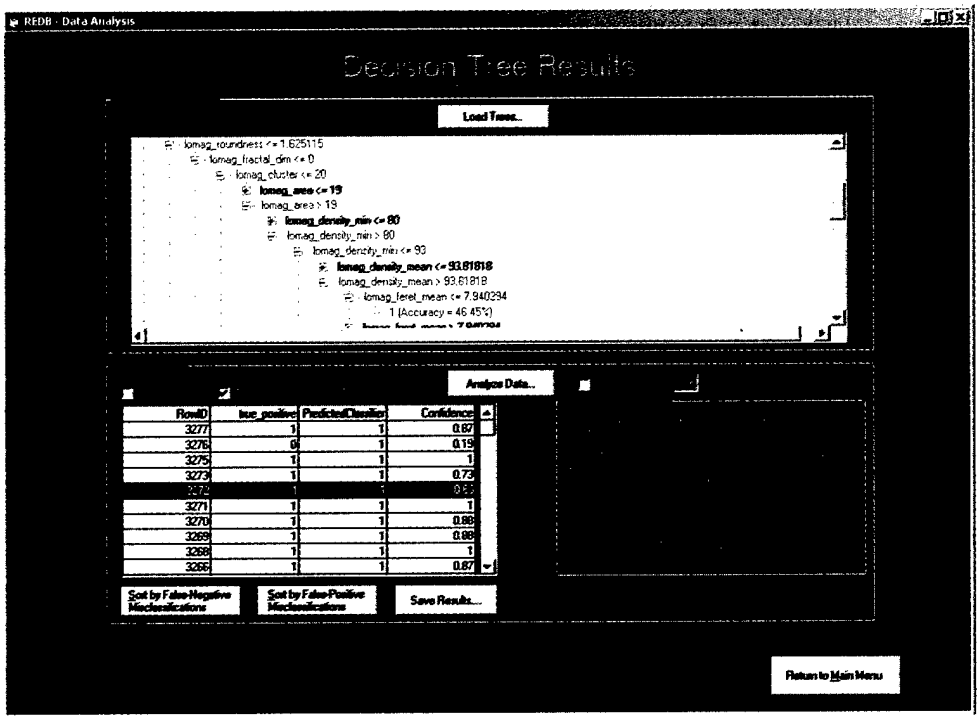
After the final set of parameters was derived, it was tested, along with the Decision Tree method described below, using the Rare Event Imaging System.

Decision Tree

The software created for the Decision Tree analysis encompassed the C5.0 algorithm and prepared the raw data from the Range Prediction method into the

format used by the algorithm. Additionally, the software served as a user interface to run the C5.0 application. Once the algorithm analyzed the raw data a tree-like structure, called a classifier, was created and served as a roadmap used to predict the outcome of new cases.

Using the algorithm, a decision tree was produced which classified 135/135 true positives, a 100% accuracy rate, and 3138/3153 false positives, a 99% accuracy rate. An image of the analysis screen is shown below.



An example of a portion of the data tree output by C5.0 is below:

```
Decision tree:

lomag_area_box <= 0.4930556: 0 (989.1)
lomag_area_box > 0.4930556:
:...lomag_density_max <= 104: 0 (237.3)
    lomag_density_max > 104:
    :...lomag_fractal_dim > 0: 0 (226.1)
        lomag_fractal_dim <= 0:
        :...lomag_diameter_mean <= 4.308567: 0 (201.9)
            lomag_diameter_mean > 4.308567:
            :...lomag_box_x_y <= 0.6266667: 0 (163.8/1.7)
                lomag_box_x_y > 0.6266667:
                :...lomag_diameter_min <= 2.242785: 1 (7.1/3.4)
                    lomag_diameter_min > 2.242785:
                    :...lomag_area_box <= 0.5649123: 0 (139.8)
                        lomag_area_box > 0.5649123:
                        :...lomag_area_box <= 0.5719298: 1 (21.4/15.5)
                            lomag_area_box > 0.5719298:
                            :...lomag_aspect > 1.968599: 0 (68.4)
                                lomag_aspect <= 1.968599:
                                :...lomag_area_box <= 0.5953947: 0 (44.2)
                                    lomag_area_box > 0.5953947:
                                    :...lomag_aspect > 1.9042: 1 (8.2/4.4)
```

```
                                    lomag_aspect <= 1.9042:
                                    :...lomag_aspect > 1.743818: 0 (66.8/0.7)
                                        lomag_aspect <= 1.743818:
                                        :...lomag_density_min > 114: 0 (68.5/2.3)
                                            lomag_density_min <= 114: [S1]

SubTree [S1]

lomag_density_min <= 80: 0 (18.6)
lomag_density_min > 80:
:...lomag_margination > 0.4966345:
    :...lomag_per_area <= 1.98e-005: 1 (16.3/7.7)
    :   lomag_per_area > 1.98e-005: 0 (16.3)
    lomag_margination <= 0.4966345:
    :...lomag_feret_mean <= 5.294395: 0 (248.5/7.7)
        lomag_feret_mean > 5.294395:
        :...lomag_feret_max > 10.198: 0 (25.7)
            lomag_feret_max <= 10.198:
            :...lomag_feret_mean <= 5.319472: 1 (10.6/5.5)
                lomag_feret_mean > 5.319472:
                :...lomag_density_max <= 123: 0 (52.9/0.7)
                    lomag_density_max > 123:
                    :...lomag_density_mean <= 106.1545: 1 (7/2)
                        lomag_density_mean > 106.1545:
                        :...lomag_diameter_max > 9.055386: 1 (16.2/6.2)
                            lomag_diameter_max <= 9.055386:
                            :...lomag_feret_max > 9.217468: 0 (18.8)
                                lomag_feret_max <= 9.217468:
                                :...lomag_perimeter > 25.46374: 1 (7.3/3)
                                    lomag_perimeter <= 25.46374:
                                    :...lomag_density_mean > 175.9275: 0 (39.3/0.7)
                                        lomag_density_mean <= 175.9275:
                                        :...lomag_margination <= 0.3354912: 1 (44.8/30.7)
                                            lomag_margination > 0.3354912:
                                            :...lomag_feret_mean > 7.82226: 0 (25.3)
                                                lomag_feret_mean <= 7.82226: [S2]

SubTree [S2]

lomag_size_length <= 6.185516: 1 (22.4/15)
lomag_size_length > 6.185516:
:...lomag_diameter_min <= 4.351791: 0 (22.3)
    lomag_diameter_min > 4.351791:
    :...lomag_feret_max <= 7.211099:
        :...lomag_density_std_dev <= 5.978093: 1 (4.2/2.7)
        :   lomag_density_std_dev > 5.978093:
        :   :...lomag_roundness > 1.315424: 0 (67/0.7)
        :       lomag_roundness <= 1.315424:
        :       :...lomag_roundness <= 1.256882: 0 (42.2/4)
        :           lomag_roundness > 1.256882: 1 (2.7/0.7)
        lomag_feret_max > 7.211099:
        :...lomag_per_area <= 2.28e-005: 1 (13.4/4.4)
            lomag_per_area > 2.28e-005:
            :...lomag_roundness > 1.407526:
                :...lomag_feret_max <= 7.810059: 0 (12.3)
                :   lomag_feret_max > 7.810059: 1 (9.2/2)
                lomag_roundness <= 1.407526:
                :...lomag_area_box <= 0.6625: 1 (16.7/9)
                    lomag_area_box > 0.6625:
                    :...lomag_aspect > 1.456438: 0 (40.6)
                        lomag_aspect <= 1.456438:
                        :...lomag_roundness > 1.367485: 0 (21.2)
                            lomag_roundness <= 1.367485:
                            :...lomag_angle <= 8.749171: 0 (19.9)
                                lomag_angle > 8.749171:
                                :...lomag_size_length <= 7.071106: 0 (13.3)
                                    lomag_size_length > 7.071106:
                                    :...lomag_feret_max <= 7.279144: 1 (5.6/2.1)
                                        lomag_feret_max > 7.279144:
                                        :...lomag_aspect > 1.416898: 1 (8.6/4.2)
                                            lomag_aspect <= 1.416898:
```

```
                                        :...lomag_roundness > 1.3667: 1 (2.1)
                                            lomag_roundness <= 1.3667: [S3]

SubTree [S3]

lomag_diameter_min > 6.991661: 1 (9.8/5)
lomag_diameter_min <= 6.991661:
:...lomag_radius_ratio > 2.421821: 1 (4.2/2.1)
    lomag_radius_ratio <= 2.421821:
    :...lomag_roundness <= 1.168165: 1 (14.6/10.1)
        lomag_roundness > 1.168165: 0 (122.7/5.2)
```

Three properties of the C5.0 algorithm were employed to optimize the probability of constructing classifiers that would accurately predict true positives from new data, they were: 1. Boosting, 2. Winnowing, and 3. Cost information.

Boosting refers to the ability of C5.0 to create multiple classifiers (called a boosting level) from one set of raw data; we used a boosting level of 10. Each one of the 10 classifiers constructed sought to eliminate those errors made by the previous classifier. In doing so, new errors were produced that served as the basis for the following classifier. After the classifiers were constructed, they were used together to predict the outcome of a new case. Specifically, each one of the 10 classifiers made a prediction on the new case along with an accuracy rate. Using a voting methodology, a final classification on the positive type, along with a confidence measurement, was returned.

Another process used by C5.0 to increase accuracy was that of winnowing. Winnowing refers to the ability of C5.0 to analyze the usefulness of all 49 parameters before it developed any classifier. Those attributes deemed to be detrimental to the final accuracy of the classifiers were unused by the algorithm. From our data, 23 parameters were winnowed for a total of 26 useful parameters.

A final step implemented to increase the true-positive identification rate was the concept of costs. By designating false-negatives as 5 times more costly than false-positives, the classifiers were constructed to maximize true-positive accuracy as opposed to overall accuracy.

RESULTS

In order to compare the Range Comparison, Decision Tree, and Combination methods, seven additional slides each from different patient samples were studied using the Rare Event Imaging System (REIS). The results of the scanning methods were compared to those of manual scanning.

Using the Range Comparison method required inputting the parameter combination chosen from the optimal parameter discovery into the REIS and scanning a slide. During the scan, each captured image was analyzed by

applying the range parameter combinations and counting those objects that fell within every range. At the end of a scan, the sum of those objects counted was referred to as the total positive count. Each positive was then reviewed manually and those objects that were actual positives were labeled as true positives, otherwise, the object was designated as a false positive.

In order to test the Decision Tree method, all restrictions on the parameters were removed and data was to be gathered from the scan without any filtering. The data gathered was then to be analyzed afterwards using the classifier returned by the Decision Tree method. However, upon attempting to gather data, it was realized that without parameters to limit the amount of data recorded by the system, it was overloaded and unable to calculate results in an efficient manner. Therefore, the focus of our testing shifted to comparing the Range Comparison and Combination methods.

Using the Combination method, the scans were run using the Range Comparison method to filter all incoming data during the scan. Before the total positives were reviewed manually, they were sent through the Decision Tree analysis. The Decision Tree analysis classified and ranked (based on confidence) all of the objects comprising the total positives, returning a list of all the objects along with their classification and confidence level. Those positives classified as true positives were then reviewed manually. The results of these experiments are shown in Table 4.

Table 4: Comparison of Range predications and combination methods

Slide No.	Manual Scan Positive Count	Range Parameter Total Positive Count	Range Parameter True Positive Count	Decision Tree True Positive Count	Decision Tree Positve Count Reduction	Decision Tree True Positive Count
1	13	36	11	12	66.67%	9
2	14	94	12	21	77.66%	11
3	0	13	0	4	69.23%	0
4	0	13	0	7	46.15%	0
5	0	49	0	7	85.71%	0
6	0	69	0	23	66.67%	0
7	0	94	0	8	91.49%	0
Average:	N/A	52.6	N/A	11.7	71.94%	N/A

As seen from the table, the average overall count of positives using the Range Parameter method alone was 52.6 positives per scan. In terms of true positives, the Range Parameter method matched all manual counts except for slide number one, where two were lost, and slide number two, where two were also lost; an overall true-positive accuracy rate of 85% was obtained. After closer inspection, it was found that all four positives lost were due to the upper range of

one parameter, the standard deviation of the density, being set too low at 60.112878. Using virtual testing, it was discovered that this upper range could be raised high enough to allow all four positives to pass without allowing any false positives to be created.

While the Range Parameter method provided encouraging results, further analyzing the data using the Decision Tree method permitted more precise numbers. Specifically, the Combination method was able to reduce the Range Parameter overall positive count per scan by an average of almost 72% (52.6 to 11.7), while still finding 86.9% of the true positives identified by the Range Parameter method. The true positives lost were likely due to the relatively small amount of data on which the decision tree classifiers were made.

The data above provides support that, between the Range Parameter method and the Combination method, the Combination method is recommended for use in the REIS.

CONCLUSION

In order to develop a system capable of quickly and accurately detecting CMV positive cells, three methods were developed to use in conjunction with the Rare Event Imaging System. Based on the resulting data, the Combination method appears to be the most successful in terms of quickly and accurately validating all potential positives.

Future plans for the CMV detection system will concentrate on improving the accuracy rate of detection. This will be accomplished by compiling a larger quantity of raw data on which to build the models. As a result of each model being a learning algorithm, the more raw data entered into the system the more accurate the resulting models will be. Another area of improvement will be to more closely analyze how the Decision Tree and Range Parameter models work cooperatively and independently of each other. Specifically, it is still of interest to use the Decision Tree model as a stand-alone screening process, as it theoretically offers speed and identification improvements over the Range Parameter method. Configuring a system that allows for its independent use is a priority for the future.

REFERENCES:

[1] van der Meer JT, Drew WL, Bowden RA, Galasso GJ, Griffiths PD, Jabs DA, Katlama C, Spector SA, Whitley RJ. Summary of the International Consensus Symposium on Advances in the Diagnosis, Treatment and Prophylaxis and Cytomegalovirus Infection. Antiviral Res 1996 Nov;32(3):119-40

[2] Kusne S, Shapiro R, Fung J. Prevention and treatment of cytomegalovirus infection in organ transplant recipients. Transpl Infect Dis. 1999 Sep;1(3):187-203. Review.

[3] Nichols WG, Boeckh M.Recent advances in the therapy and prevention of CMV infections. J Clin Virol. 2000 Feb;16(1):25-40.

[4] Boeckh M, Gooley TA, Myerson D, Cunningham T, Schoch G, Bowden RA. Cytomegalovirus pp65 antigenemia-guided early treatment with ganciclovir versus ganciclovir at engraftment after allogeneic marrow transplantation: a randomized double-blind study. Blood. 1996 Nov 15;88(10):4063-71.

[5] Sharma AK, Taylor JD, Tong W, Brown MW, Sells RA, Bakran A. Utility of the pp65 direct antigenemia test in the diagnosis of cytomegalovirus (CMV) in renal transplant recipients. Transplant Proc. 1997 Feb-Mar;29(1-2):799.

[6] Kraeft SK, Sutherland R, Gravelin L, Hu GH, Ferland LH, Richardson P, Elias A, Chen LB.Detection and analysis of cancer cells in blood and bone marrow using a rare event imaging system. Clin Cancer Res. 2000 Feb;6(2):434-42.

[7] Kraeft SK, et al. Reliable and Sensitive Identification of Occult Tumor Cells Using the Improved Rare Event Imaging System. Manuscript in preparation

[8] Quinlan, J. R. *C4.5: Programs for Machine Learning.* Morgan Kaufmann. San Mateo, CA, 1993.